

Relative entropy via non-sequential recursive pair substitution

D. Benedetto¹, E. Caglioti¹, G. Cristadoro², M. Degli Esposti²

¹ Dipartimento di Matematica, Università di Roma “La Sapienza”, P.le A. Moro 5, 00185 Roma, Italy. e-mail: benedetto@mat.uniroma1.it ; caglioti@mat.uniroma1.it

² Dipartimento di Matematica, Università di Bologna, P.za di Porta San Donato 5, 40126 Bologna, Italy. e-mail: desposti@dm.unibo.it ; cristadoro@dm.unibo.it

Abstract. The entropy of an ergodic source is the limit of properly rescaled 1-block entropies of sources obtained applying successive non-sequential recursive pairs substitutions [7],[2]. In this paper we prove that the cross entropy and the Kullback-Leibler divergence can be obtained in a similar way.

AMS classification scheme numbers: 94A17

Keywords: information theory, source and channel coding, relative entropy.

1. Introduction

Kullback-Leibler (KL) divergence (relative entropy) can be considered as a measure of the difference/dissimilarity between sources. Estimating KL divergence from finite realizations of a stochastic process with unknown memory is a long-standing problem, with interesting mathematical aspects and useful applications to automatic categorization of symbolic sequences. Namely, an empirical estimation of the divergence can be used to classify sequences (for approaches to this problem using other methods, in particular true metric distances, see [10], [12]; see also [1]).

In [16] Ziv and Merhav showed how to estimate the KL divergence between two sources, using the parsing scheme of LZ77 algorithm [15] on two finite length realizations. They proved the consistence of the method by showing that the estimate of the divergence for two markovian sources converges to their relative entropy when the length of the sequences diverges. Furthermore they proposed this estimator as a tool for an “universal classification” of sequences.

A procedure based on the implementations of LZ77 algorithm (gzip, winzip) is proposed in [3]. The estimate obtained of the relative entropy is then used to construct phylogenetic trees for languages and is proposed as a tool to solve authorship attribution problems. Moreover, the relation between the relative entropy and the estimate given by this procedure is analyzed in [13].

Two different algorithms are proposed and analyzed in [5], see also [6]. The first one is based on the Burrows-Wheeler block sorting transform [4], while the other

uses the Context Tree Weighting method. The authors proved the consistence of these approximation methods and show that these methods outperform the others in experiments.

In [2] it is shown how to construct an entropy estimator for stationary ergodic stochastic sources using non-sequential recursive pairs substitutions method, introduced in [7] (see also [9] and references therein for similar approaches).

In this paper we want to discuss the use of similar techniques to construct an estimator of relative (and cross) entropy between a pair of stochastic sources. In particular we investigate how the asymptotic properties of concurrent pair substitutions might be used to construct an optimal (in the sense of convergence) relative entropy estimator. A second relevant question arises about the computational efficiency of the derived indicator. While here we address the first, mostly mathematical, question, we leave the computational and applicative aspects for forthcoming research.

The paper is structured as follows: in section 2 we state the notations, in section 3 we describe the details of the non-sequential recursive pair substitutions (NSRPS) method, in section 4 we prove that NSRPS preserve the cross and the relative entropy, in section 5 we prove the main result: we can obtain an estimate of the relative entropy by calculating the 1-block relative entropy of the sequences we obtain using the NSRPS method.

2. Definitions and notations

We introduce here the main definitions and notations, often following the formalism used in [2]. Given a finite alphabet \mathcal{A} , we denote with $\mathcal{A}^* = \cup_{k \geq 1} \mathcal{A}^k$ the set of finite words. Given a word $\omega \in \mathcal{A}^n$, we denote by $|\omega| = n$ its length and if $1 \leq i < j \leq n$ and $\omega = (\omega_1, \omega_2, \dots, \omega_n)$, we use ω_i^j to indicate the subword $\omega_i^j = (\omega_i, \dots, \omega_j)$. We use similar notations for one-sided infinite (elements of $\mathcal{A}^{\mathbb{N}}$) or double infinite words (elements of $\mathcal{A}^{\mathbb{Z}}$). Often sequences will be seen as finite or infinite realizations of discrete-time stochastic stationary, ergodic processes of a random variable X with values in \mathcal{A} .

The n -th order joint distributions μ_n identify the process and its elements follow the consistency conditions:

$$\mu_n(\omega_1^n) = \sum_{\omega_0 \in \mathcal{A}} \mu_{n+1}(\omega_0, \omega_1, \dots, \omega_n) = \sum_{\omega_{n+1} \in \mathcal{A}} \mu_{n+1}(\omega_1, \dots, \omega_n, \omega_{n+1}).$$

When no confusion will arise, the subscript n will be omitted, and we will just use $\mu(\omega_1^n)$ to denote both the measure of the cylinder and the probability of the finite word.

Equivalently, a distribution of a process can also be defined by specifying the initial one-character distribution μ_1 and the successive conditional distributions:

$$\mu(\omega_n | \omega_1^{n-1}) = \frac{\mu_n(\omega_1^n)}{\mu_{n-1}(\omega_1^{n-1})}.$$

Given an ergodic, stationary stochastic source we define as usual:

n-block entropy

$$H_n(\mu) := - \sum_{|\omega|=n} \mu(\omega) \log \mu(\omega).$$

n-conditional entropy

$$h_n(\mu) := H_{n+1}(\mu) - H_n(\mu) = \sum_{\omega_1^n \in \mathcal{A}^n, a \in \mathcal{A}} \mu(\omega_1^n a) \log \mu(a | \omega_1^n) := E_{\mu_{n+1}}(\log \mu(a | \omega_1^n)),$$

where $\omega_1^n a$ denotes the concatenated word $(\omega_1, \omega_2, \dots, \omega_n, a)$ and $E_\mu(\bullet)$ is just the process average.

Entropy of μ

$$h(\mu) := \lim_{n \rightarrow \infty} \frac{H_n(\mu)}{n} = \lim_{n \rightarrow \infty} h_n(\mu) = E_\mu(\log \mu(a | \omega_1^\infty))$$

The following properties and results are very well known [14], but at the same time quite important for the proofs and the techniques developed here (and also in [2]):

- $h(\mu) \leq \dots \leq h_k(\mu) \leq h_{k-1}(\mu) \leq \dots \leq h_1(\mu) \leq H_1(\mu)$.
- A process μ is k -Markov if and only if $h(\mu) = h_k(\mu)$.
- *Entropy Theorem*: for almost all realizations of the process, we have

$$h(\mu) = \lim_{n \rightarrow \infty} -\frac{1}{n} \log \mu(x_1^n), \quad \mu - \text{a.s.}$$

In this paper we focus on properties involving pairs of stochastic sources on the same alphabet with distributions μ and ν , namely *cross entropy* and the related *relative entropy* (or *Kullback-Leibler divergence*):

n-conditional cross entropy

$$h_n(\mu || \nu) = - \sum_{\omega \in \mathcal{A}^n, a \in \mathcal{A}} \mu(\omega a) \log \nu(a | \omega), \quad (2.1)$$

cross entropy

$$h(\mu || \nu) = \lim_{n \rightarrow +\infty} h_n(\mu || \nu), \quad (2.2)$$

relative entropy (Kullback-Leibler divergence)

$$\begin{aligned} d(\mu || \nu) &= \lim_{n \rightarrow \infty} E_\mu \left(\log \frac{\mu(\omega_n | \omega_1^{n-1})}{\nu(\omega_n | \omega_1^{n-1})} \right) \\ &= \lim_{n \rightarrow \infty} \sum_{\omega_1^n \in \mathcal{A}^n} \mu(\omega_1^n) \log \frac{\mu(\omega_n | \omega_1^{n-1})}{\nu(\omega_n | \omega_1^{n-1})}. \end{aligned} \quad (2.3)$$

Note that

$$h(\mu || \nu) = h(\mu) + d(\mu || \nu)$$

Moreover we stress that, if ν is k -Markov then, for any μ

$$h(\mu || \nu) = h_k(\mu || \nu) \quad (2.4)$$

Namely $h_l(\mu\|\nu) = h_k(\mu\|\nu)$ for any $l \geq k$:

$$\begin{aligned} h_l(\mu\|\nu) &= -\sum_{\omega \in A^l, a \in A} \mu(\omega a) \log \nu(a|\omega) = -\sum_{\omega \in A^{l-k}, b \in A^k, a \in A} \mu(\omega ba) \log \nu(a|\omega b) \\ &= -\sum_{\omega \in A^{l-k}, b \in A^k, a \in A} \mu(\omega ba) \log \nu(a|b) \\ &= -\sum_{b \in A^k, a \in A} \mu(ba) \log \nu(a|b) = h_k(\mu\|\nu) \end{aligned}$$

Note that $h_1(\mu\|\nu)$ depends only on the two-symbol distribution of μ .

Entropy and cross entropy can be related to the asymptotic behavior of properly defined *returning times* and *waiting times*, respectively. More precisely, given an ergodic, stationary process μ , a sample sequence $w = w_1, w_2, \dots$ and $n \geq 1$, we define the returning time of the first n characters as:

$$\text{returning time} \quad R(w_1^n) = \min\{k > 1 : w_k^{k+n-1} = w_1^n\} \quad (2.5)$$

Similarly, given two realizations $w = w_1, \dots, w_n, \dots$ and $z = z_1, \dots, z_n, \dots$ of μ and ν respectively, we define the

$$\text{waiting time} \quad W(w_1^n, z) = \min\{k > 1 : z_k^{k+n-1} = w_1^n\} \quad (2.6)$$

Obviously $W(w_1^n, w) = R(w_1^n)$.

We now have the following two important results:

Theorem 2.1 (Entropy and returning time [11]) *If μ is a stationary, ergodic process, then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log R(w_1^n) = h(\mu) \quad \mu\text{-a.s.}$$

Theorem 2.2 (Relative entropy and waiting time [8]) *If μ is stationary and ergodic, ν is k -Markov and the marginals μ_n of μ are dominated by the corresponding marginals ν_n of ν , i.e. $\mu_n \ll \nu_n$, then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log W(w_1^n, z) = h(\mu) + d(\mu\|\nu) = h(\mu\|\nu), \quad (\mu \times \nu)\text{-a.s.}$$

3. Non sequential recursive pair substitutions

We now introduce a family of transformations on sequences and the corresponding operators on distributions: given $a, b \in \mathcal{A}$ (including $a = b$), $\alpha \notin \mathcal{A}$ and $\mathcal{A}' = \mathcal{A} \cup \{\alpha\}$, a *pair substitution* is a map $G_{ab}^\alpha : \mathcal{A}^* \rightarrow \mathcal{A}'^*$ which substitutes sequentially, from left to right, the occurrences of ab with α . For example

$$G_{01}^2(0010001011100100) = 020022110200.$$

or:

$$G_{00}^2(0001000011) = 2012211.$$

$G = G_{ab}^\alpha$ is always an injective but not surjective map that can be immediately extended also to infinite sequences $w \in \mathcal{A}^\mathbb{N}$.

The action of G shorten the original sequence: we denote by Z the inverse of the contraction rate:

$$\frac{1}{Z_{ab}(\omega_1^n)} := \frac{|G_{ab}^\alpha(\omega_1^n)|}{|\omega_1^n|} = 1 - \frac{\#\{ab \subseteq \omega_1^n\}}{n}$$

For μ -typical sequences we can pass to the limit and define:

$$\frac{1}{Z^\mu} := \lim_{n \rightarrow \infty} \frac{|G(\omega_1^n)|}{|\omega_1^n|} = \begin{cases} 1 - \mu(ab) & \text{if } a \neq b \\ 1 - \mu(aa) + \mu(aaa) - \mu(aaaa) + \dots & \text{if } a = b \end{cases}$$

An important remark is that if we start from a source where admissible words are described by constraints on consecutive symbols, this property will remain true even after an arbitrary pair substitution. In other words (see Theorem 2.1 in [2]): a pair substitution maps pair constraints in pair constraints.

A pair substitution G_{ab}^α naturally induces a map on the set of ergodic stationary measures on $A^\mathbb{Z}$ by mapping typical sequences w.r.t. the original measure μ in typical sequences w.r.t. the transformed measure $\mathcal{G}\mu$: given $z_1^k \in \mathcal{A}^*$ then (Theorem 2.2 in [2])

$$\mathcal{G}\mu(z_1^k) := \lim_{n \rightarrow \infty} \frac{\#\{z_1^k \subseteq G(\omega_1^n)\}}{|G(\omega_1^n)|}$$

exists and is constant μ almost everywhere in $\omega \in \mathcal{A}^\mathbb{N}$, moreover $\{\mathcal{G}\mu(z)\}_{z \in \mathcal{A}^*}$ are the marginals of an ergodic measure on $\mathcal{A}^\mathbb{Z}$.

Again in [2], the following results are proved showing how entropies transform under the action of $\mathcal{G} = \mathcal{G}_{ab}^\alpha$, with expanding factor $Z = Z_{ab}^\mu$:

Invariance of entropy

$$h(\mathcal{G}\mu) = Z h(\mu).$$

Decreasing of the 1-conditional entropy

$$h_1(\mathcal{G}\mu) \leq Z h_1(\mu).$$

Moreover, \mathcal{G} maps 1-Markov measures in 1-Markov measures. In fact:

$$h(\mathcal{G}\mu) \leq h_1(\mathcal{G}\mu) \leq Z h_1(\mu) = Z h(\mu) = h(\mathcal{G}\mu)$$

Decreasing of the k -conditional entropy

$$h_k(\mathcal{G}\mu) \leq Z h_k(\mu).$$

Moreover \mathcal{G} maps k -Markov measures in k -Markov measures.

While later on we will give another proof of the first fact, we remark that this property, together with the decrease of the 1-conditional entropy, reflect, roughly speaking, the fact that the amount of information of $G(\omega)$, which is equal to that of ω , is more concentrated on the pairs of consecutive symbols.

As we are interested in sequences of recursive pair substitutions, we assume to start with an initial alphabet \mathcal{A} and define an increasing alphabet sequence $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_N, \dots$. Given N and chosen $a_N, b_N \in \mathcal{A}_{N-1}$ (not necessarily different):

we indicate with $\alpha_N \notin A_{N-1}$ a new symbol and define the new alphabet as $\mathcal{A}_N = \mathcal{A}_{N-1} \cup \{\alpha_N\}$;

we denote with G_N the substitution map $G_N = G_{a_N b_N}^{\alpha_N} : \mathcal{A}_{N-1}^* \rightarrow \mathcal{A}_N^*$ which substitutes whit α_N the occurrences of the pair $a_N b_N$ in the strings on the alphabet \mathcal{A}_{N-1} ;

we denote with \mathcal{G}_N the corresponding map from the measures on $A_{N-1}^{\mathbb{Z}}$ to the measures on $A_N^{\mathbb{Z}}$;

we define by Z_N the corresponding normalization factor $Z_N = Z_{a_N b_N}^{\alpha_N}$.

We use the over-line to denote iterated quantities:

$$\bar{G}_N := G_N \circ G_{N-1} \circ \cdots \circ G_1, \quad \bar{\mathcal{G}}_N := \mathcal{G}_N \circ \mathcal{G}_{N-1} \circ \cdots \circ \mathcal{G}_1$$

and also

$$\bar{Z}_N = Z_N Z_{N-1} \cdots Z_1.$$

The asymptotic properties of \bar{Z}_N clearly depend on the pairs chosen in the substitutions. In particular, if at any step N the chosen pair $a_N b_N$ is the pair of maximum of frequency of \mathcal{A}_{N-1} then (Theorem 4.1 in [2]):

$$\lim_{N \rightarrow \infty} \bar{Z}_N = +\infty$$

Regarding the asymptotic properties of the entropy we have the following theorem that rigorously show that $\mu_N := \bar{G}_N \mu$ becomes asymptotically 1-Markov:

Theorem 3.1 (Entropy via NSRPS [2]) *If*

$$\lim_{N \rightarrow \infty} \bar{Z}_N = +\infty$$

then

$$h(\mu) = \lim_{N \rightarrow \infty} \frac{1}{\bar{Z}_N} h_1(\mu_N)$$

The main results of this paper is the generalization of this theorem to the cross and relative entropy.

Before entering in the details of our construction let us sketch here the main steps.

In particular let us consider the cross entropy (the same argument will apply to the relative entropy) of the measure μ with respect to the measure ν : i.e. $h(\mu||\nu)$.

As we will show, but for the normalization factor \bar{Z}_N^μ , this is equal to the cross entropy of the measure $G_N \mu$ w.r.t the measure $G_N \nu$:

$$h(\mu||\nu) = \frac{h(G_N \mu || G_N \nu)}{\bar{Z}_N^\mu}$$

Moreover, as we have seen above, if we choose the substitution in a suitable way (for instance if at any step we substitute the pair with maximum frequency) then $\bar{Z}_N^\nu \rightarrow \infty$ and the measure $G_N \nu$ becomes asymptotically 1-Markov as $N \rightarrow \infty$.

Interestingly, we do not know if Z_ν also diverges (we will discuss this point in the sequel).

Nevertheless, noticing that the cross entropy of a 1-Markov source w.r.t a generic ergodic source is equal to the 1-Markov cross entropy between the two sources, it is reasonable to expect that the cross entropy $h(\mu||\nu)$ can be obtained as the following limit:

$$h(\mu||\nu) = \lim_{N \rightarrow +\infty} \frac{h_1(\mathcal{G}_N \mu || \mathcal{G}_N \nu)}{\overline{Z}_N^\mu}$$

This is exactly what we will prove in the two next sections.

4. Scaling of (relative) entropy via waiting times

We first show how the relative entropy between two stochastic process μ and ν scales after acting with the *same* pair substitution on both sources to have $\mathcal{G}\mu$ and $\mathcal{G}\nu$. More precisely we make use of Theorem 2.2 and have the following:

Theorem 4.1 (Invariance of relative entropy for pair substitution) *If μ is ergodic, ν is a Markov chain and $\mu_n \ll \nu_n$, then if G is a pair substitution*

$$d(\mathcal{G}\mu || \mathcal{G}\nu) = Z^\mu d(\mu || \nu)$$

Proof. To fix the notations, let us denote by w and z the infinite realizations of the process of measure μ and ν respectively, and by w_k^n and z_k^n the corresponding finite substrings. Let us denote by $a, b \in \mathcal{A}$ the characters involved in the pair substitution $G = G_{ab}^\alpha$. Moreover let us denote the waiting time with the shorter notation:

$$t_n := W(w_1^n, z).$$

We now explore how the waiting time rescale with respect to the transformation G : we consider the first time we see the sequence $G(w_1^n)$ inside the sequence $G(z)$. To start with, we assume that $w_1 \neq b$ as we can always consider Th. 2.2 for realizations with a fixed prefix of positive probability. Moreover we choose a subsequence $\{n_i\}$ such that n_i is the smallest $n > n_{i-1}$ such that $w_{n_i} \neq a$. Of course $n_i \rightarrow \infty$ as $i \rightarrow \infty$. In this case, it is easy to observe that

$$W(G(w_1^{n_i}), G(z)) = |G(w_1^{t_{n_i}})|$$

Then, using Theorem 2.2

$$\begin{aligned} h(\mathcal{G}\mu || \mathcal{G}\nu) &= \lim_{i \rightarrow +\infty} \frac{1}{|G(w_1^{n_i})|} \log [W(G(w_1^{n_i}), G(z))] = \\ &= \lim_{i \rightarrow +\infty} \frac{n_i}{|G(w_1^{n_i})|} \frac{\log |G(w_1^{t_{n_i}})|}{n_i} = \\ &= \lim_{i \rightarrow +\infty} \frac{n_i}{|G(w_1^{n_i})|} \left[\frac{1}{n_i} \log(t_{n_i}) + \frac{1}{n_i} \log \left(\frac{|G(w_1^{t_{n_i}})|}{t_{n_i}} \right) \right] = \\ &= Z^\mu h(\mu || \nu) \end{aligned} \tag{4.7}$$

where in the last step we used the fact that $t_{n_i} \rightarrow \infty$ as $i \rightarrow \infty$, the definition of Z^μ and Theorem 2.2 for μ and ν . Note that for $\mu = \nu$, equation (4.7) reproduces the content of Theorem 3.1 of [2]:

$$h(\mathcal{G}\mu) = Z^\mu h(\mu),$$

that thus implies

$$d(\mathcal{G}\mu || \mathcal{G}\nu) = Z^\mu d(\mu || \nu).$$

Note that the limit in Th. 2.2 is almost surely unique and then the initial restrictive assumption $w_1 \neq b$ and the use of the subsequence n_i have no consequences on the thesis; this concludes the proof.

□

Before discussing the convergence of relative entropy under successive substitutions we go thorough a simple explicit example of the Theorem 4.1, in order to show the difficulties we deal with, when we try to use the explicit expressions of the transformed measures we find in [2].

Example. We treat here the most simple case: μ and ν are Bernoulli binary processes with parameters $\mu(0), \mu(1)$ and $\nu(0), \nu(1)$ respectively. We consider the substitution $G = G_{01}^2$ given by $01 \rightarrow 2$. It is long but easy to verify that $\mathcal{G}\mu$ is a stationary, ergodic, 1-Markov with equilibrium state

$$\mathcal{G}\mu(0) = Z\mu(00), \quad \mathcal{G}\mu(1) = Z(\mu(1) - \mu(01)), \mathcal{G}\mu(2) = Z\mu(01) ,$$

where $Z = Z^\mu(01) = (1 - \mu(01))^{-1}$.

For example, given a $\mathcal{G}\mu$ -generic sequence y_1, \dots, y_m , corresponding to a μ -generic sequence x_1, \dots, x_n ($y = Gx$):

$$\begin{aligned} \mathcal{G}\mu(0) &= \lim_{m \rightarrow \infty} \frac{1}{m} \# \{0 \in y_1^m\} \\ &= \lim_{n \rightarrow \infty} \frac{n}{m} \cdot \frac{\# \{0 \in x_1^n\} - \# \{01 \in x_1^n\}}{n} \\ &= (\mu(0) - \mu(01)) \cdot \lim_{n \rightarrow \infty} \frac{n}{n - \# \{01 \in x_1^n\}} = Z(\mu(0) - \mu(01)) = Z\mu(00) \end{aligned}$$

Clearly:

$$\mathcal{G}\mu(0) + \mathcal{G}\mu(1) + \mathcal{G}\mu(2) = 1$$

Using the same argument as before, it is now possible to write down the probability distribution of pair of characters for $\mathcal{G}\mu$. Again the following holds for a generic process:

$$\begin{array}{lll} \frac{\mathcal{G}\mu(00)}{Z} = \mu(00) - \mu(001) = \mu(000) & \frac{\mathcal{G}\mu(01)}{Z} = 0 & \frac{\mathcal{G}\mu(02)}{Z} = \mu(001) \\ \frac{\mathcal{G}\mu(10)}{Z} = \mu(10) - \mu(010) - \mu(101) + \mu(0101) & \frac{\mathcal{G}\mu(11)}{Z} = \mu(11) - \mu(011) & \frac{\mathcal{G}\mu(12)}{Z} = \mu(101) - \mu(0101) \\ \frac{\mathcal{G}\mu(20)}{Z} = \mu(010) - \mu(0101) & \frac{\mathcal{G}\mu(21)}{Z} = \mu(011) & \frac{\mathcal{G}\mu(22)}{Z} = \mu(0101) \end{array}$$

It is easy to see that $\sum_{x,y=0,1,2} \mathcal{G}\mu(xy) = 1$. Now we can write the transition matrix P for the process $\mathcal{G}\mu$ as $P(y|x) = \mathcal{G}\mu(xy)/\mathcal{G}\mu(x)$:

$$P = \begin{pmatrix} P(0|0) & P(1|0) & P(2|0) \\ P(0|1) & P(1|1) & P(2|1) \\ P(0|2) & P(1|2) & P(2|2) \end{pmatrix}$$

For Bernoulli processes:

$$P = \begin{pmatrix} \mu(0) & 0 & \mu(1) \\ \mu(00) & \mu(1) & \mu(01) \\ \mu(00) & \mu(1) & \mu(01) \end{pmatrix}.$$

We now denote with Q the transition matrix for $Q\nu$. For the two 1-Markov processes, we have

$$d(\mathcal{G}\mu||\mathcal{G}\nu) = \sum_{x=0,1,2} \mu(x) \sum_{y=0,1,2} P(y|x) \log \frac{P(y|x)}{Q(y|x)}.$$

Via straightforward calculations, using the product structure of the measure μ :

$$\begin{aligned} d(G\mu||G\nu) &= Z\mu(00) \left[\mu(0) \log \frac{\mu(0)}{\nu(0)} + \mu(1) \log \frac{\mu(1)}{\nu(1)} \right] \\ &+ Z\mu(11) \left[\mu(00) \log \frac{\mu(00)}{\nu(00)} + \mu(1) \log \frac{\mu(1)}{\nu(1)} + \mu(01) \log \frac{\mu(01)}{\nu(01)} \right] \\ &+ Z\mu(01) \left[\mu(00) \log \frac{\mu(00)}{\nu(00)} + \mu(1) \log \frac{\mu(1)}{\nu(1)} + \mu(01) \log \frac{\mu(01)}{\nu(01)} \right] \\ &= Z\mu(00)d(\mu||\nu) + Z\mu(1) \left[\mu(00) \log \frac{\mu(00)}{\nu(00)} + \mu(1) \log \frac{\mu(1)}{\nu(1)} + \mu(01) \log \frac{\mu(01)}{\nu(01)} \right] \\ &= Z\mu(00)d(\mu||\nu) + Z\mu(1) [\mu(0)d(\mu||\nu) + d(\mu||\nu)] \\ &= Zd(\mu||\nu)(\mu(00) + \mu(10) + \mu(1)) \\ &= Zd(\mu||\nu) \end{aligned}$$

5. The convergence

We now prove that the renormalized 1-Markov cross entropy between μ_n and ν_n converges to the cross-entropy between $G_n\mu$ and $G_n\nu$ as the number of pair substitution n goes to ∞ .

More precisely:

Theorem 5.1 (KL divergence via NSRPS) *If $\bar{Z}_N^\nu \rightarrow +\infty$ as $N \rightarrow +\infty$,*

$$h(\mu||\nu) = \lim_{N \rightarrow +\infty} \frac{h_1(\mathcal{G}_N\mu||\mathcal{G}_N\nu)}{\bar{Z}_N^\mu}$$

Proof. Let us define, as in [2] the following operators on the ergodic measures: \mathcal{P} is the projection operator that maps a measure to its 1-Markov approximation, whereas \mathcal{P}_N is the operator such that for any arbitrary ν

$$\mathcal{P}\bar{\mathcal{G}}_N\nu = \bar{\mathcal{G}}_N\mathcal{P}_N\nu$$

We notice (see [2] for the details) that the normalization constant for $\mathcal{P}_N\nu$ is the same of that for ν :

$$Z_N^\nu = Z_N^{\mathcal{P}_N\nu}.$$

The measure $\mathcal{P}_N \nu$ is not 1-Markov, but we know that it becomes 1-Markov after N steps of substitutions, in fact it becomes $\mathcal{P}\bar{G}_N \nu$. Moreover, as discussed in [2], it is an approximation of ν if \bar{Z}_N^ν diverges: for any ω of length k ,

$$|\mathcal{P}_N \nu(\omega) - \nu(\omega)| \leq 2 \frac{k^2}{\bar{Z}_N^\nu} \quad (5.8)$$

Now it is easy to establish the following chain of equalities:

$$h(\mu || \mathcal{P}_N \nu) = \frac{1}{\bar{Z}_N^\mu} h(\mathcal{G}_N \mu || \mathcal{G}_N \mathcal{P}_N \nu) = \frac{1}{\bar{Z}_N^\mu} h(\mathcal{G}_N \mu || \mathcal{P}\mathcal{G}_N \nu) = \frac{1}{\bar{Z}_N^\mu} h_1(\mathcal{G}_N \mu || \mathcal{G}_N \nu)$$

where we have used the conservation of the cross entropy h and the fact that $H(\pi || \xi) = h_1(\pi || \xi)$ if ξ are 1-Markov, as shown in eq. 2.4. To conclude the proof we have to show that

$$h(\mu || \mathcal{P}_N \nu) \rightarrow h(\mu || \nu)$$

This is an easy consequence of eq. 5.8 the definition 2.1 and eq. 2.2. \square

6. Conclusions and remarks

It is important to remark that we are not assuming the divergence of \bar{Z}_N^μ too, as not being necessary for the convergence to the (rescaled) two-characters relative entropy.

Nevertheless, it would be interesting to understand both the topological and statistical constraints that prevent or permit the divergence of the expanding factor \bar{Z}_N^μ . Experimentally, it seems that if we start with two measures with finite relative entropy (i.e. with absolutely continuous marginals), then if we choose the standard strategy (most frequent pair substitution) for the sequence of pair substitutions that yields the divergence of \bar{Z}_N^ν , we also simultaneously obtain the divergence of \bar{Z}_N^μ (see for instance fig. 1).

On the other hand, it seems possible to consider particular sources and particular strategies of pairs substitutions withdiverging \bar{Z}_N^ν , that prevent the divergence of \bar{Z}_N^μ . At this moment we do not have conclusive rigorous mathematical results on this subject.

Finally, let us note that Th. 5.1 do not give directly an algorithm to estimate the relative entropy: in any implementation we would have to specify the “optimal” number of pairs substitutions, with respect to the length of initial sequences and also with respect to the dimension of the initial alphabet. Namely, in the estimate we have to take into account at least two correction terms, which diverges with N : the entropy cost of writing the substitutions and the entropy cost of writing the frequencies of the pairs of characters in the alphabet we obtain after the substitutions (or equivalent quantities if we use, for instance, arithmetic codings modeling the two character frequencies).

For what concerns possible implementations of the method it is important to notice that the NSRPS procedure can be implemented in linear time [9]. Therefore it seems reasonable that reasonably fast algorithms to compute relative entropy via NSRPS can be designed. Anyway, preliminary numerical experiments show that for sources of

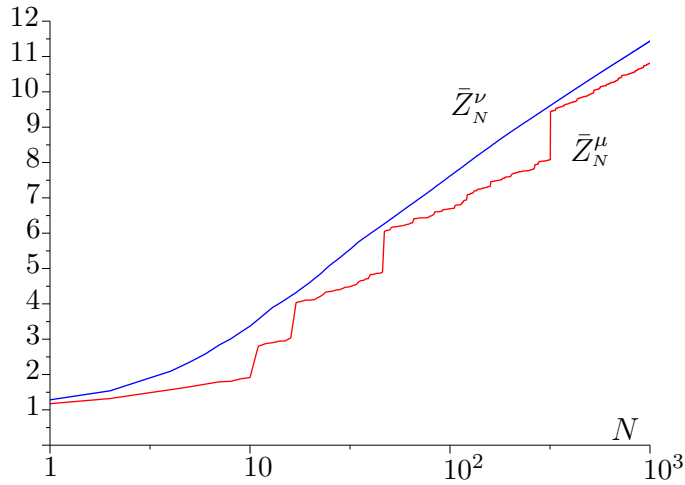


Figure 1. \bar{Z}_N^ν and \bar{Z}_N^μ for realizations of 10^8 characters of two Markov process on $\{0,1\}$ with memory of length 5. The pairs are chosen with the standard strategy (most frequent pair substitution) for the sequences of starting measure ν .

finite memory this method seems to have the same limitations of that based on parsing procedures, with respect to the methods based on the analysis of context introduced in [5].

In fig. 2 we show the convergence of the estimates of the entropies of the two sources and of the cross entropy, given Th. 5.1, for two Markov process of memory 5. In this case, the numbers of substitutions $N = 20$ is small with respect to the length of the sequences 10^8 , then the correction terms are negligible.

Let us finally note that the cross entropy estimate might show large variations for particular values of N . This could be interpreted by the fact that for these values of N pairs with particular relevance for one source with respect to the other have been substituted. This example suggest that the NSRPS method for the estimation of the cross entropy should be useful in sequences analysis, for example in order to detect strings with a peculiar statistical role.

References

- [1] C. Basile, D. Benedetto, E. Caglioti, M. Degli Esposti: An example of mathematical authorship attribution. *J. Math. Phys.* **41** 125211 (2008) doi:10.1063/1.29965079
- [2] D. Benedetto, E. Caglioti, D. Gabrielli: Non-sequential recursive pair substitution: some rigorous results. *Jour. Stat. Mech. Theo. Exp.* ISSN: 1742-5468 (on line) **09** pp. 1–21 doi:10.1088/1742.-5468/2006/09/P09011 (2006)
- [3] D. Benedetto, E. Caglioti, V. Loreto: Language Trees and Zipping, *Physical Review Letters* **88** 4 (2002)
- [4] Burrows, M., and Wheeler D. J.: A block sorting lossless data compression algorithm. *Tech. Rep. 124. Digital Equipment Corporation, Palo Alto, Calif.* (1994).

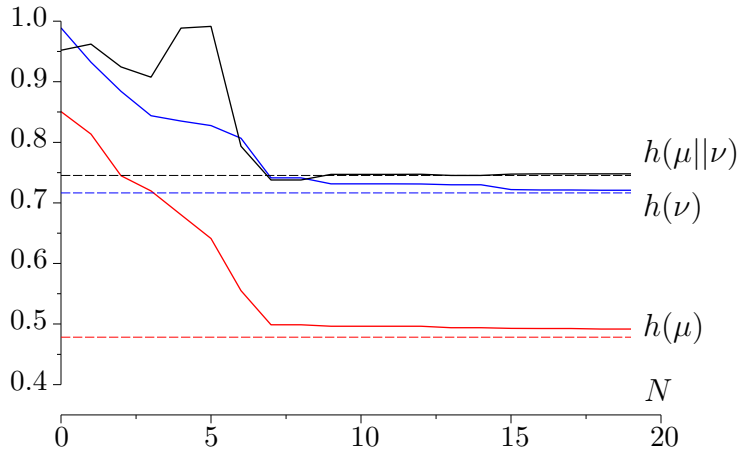


Figure 2. Solid lines are the estimates of $h(\nu)$, $h(\mu)$ and of the cross entropy $h(\mu|\nu)$ obtained after N pairs substitutions. The dashed lines are the corresponding analytic value.

- [5] H. Cai, S. R. Kulkarni, S. Verdù: Universal Divergence Estimation for Finite-Alphabet Sources. *IEEE Trans. inf. theory* **52** no. 8 (2006)
- [6] H. Cai, S. Kulkarni and S. Verdù: Universal Estimation of Entropy via Block Sorting, *IEEE Trans. Information Theory*, **50**, no. 7 (2004).
- [7] P. Grassberger: Data compression and entropy estimates by non-sequential recursive pair substitution, *ArXiv:physics/0207023*
- [8] I. Kontoyiannis: Asymptotic recurrence and waiting times for stationary processes. *Journal of Thoretical Probability* **11** (1998), pp. 795-811.
- [9] N. Jesper Larsson and A. Moffat: Off-Line Dictionary Based Compression. *IEEE Transactions on Information Theory* **88** no. 11 pp. 1722-1732 (2000).
- [10] M. Li, X. Chen, X. Li, B. Ma, P. Vitanyi: The similarity metric. *IEEE Trans. Inf. Theory* **50** no. 12 pp. 3250-3264 (2004).
- [11] D.S. Ornstein, B.Weiss: Entropy and data compression schemes. *IEEE Transactions on Information Theory* **39**, 1 (1993), pp. 78-83.
- [12] H.H. Otu, K. Sayood: A new sequence distance measure for phylogenetic tree construction. *Bioinformatics* **19**, 16 (2003)
- [13] A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, A. Vulpiani: Data Compression and Learning in Time Sequences Analysis. *Pysica D* **180** pp. 92-107 (2003)
- [14] P.C. Shields: The Ergodic Theory of Discrete Sample Paths. *American Mathematical Society* (1996).
- [15] J. Ziv, A. Lempel: A Universal Algorithm for Sequential Data Compression. *IEEE Transactions on Information Theory* **23**, 3 (1977), pp. 337-343.
- [16] J. Ziv, N. Merhav: A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory* **39**, 4 (1993), pp. 1270-1279.